GENETICS

doi: 10.3389/fgene.2013.00067

Detection and impact of rare regulatory variants in human disease

Xin Li [1,2] **** and Stephen B. Montgomery*** [1,2] ******

[1] *Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA*
[2] *Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA*

Edited by:
Liming Liang, Harvard School of Public Health, USA

Reviewed by:
Yuehua Cui, Michigan State University, USA
Jasmin Divers, Wake Forest University, USA

****Correspondence:***
Xin Li and Stephen B. Montgomery, Department of Pathology, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305-5324, USA.
e-mail: xli6@stanford.edu; smontgom@stanford.edu

Advances in genome sequencing are providing unprecedented resolution of rare and private variants. However, methods which assess the effect of these variants have relied predominantly on information within coding sequences. Assessing their impact in non-coding sequences remains a significant contemporary challenge. In this review, we highlight the role of regulatory variation as causative agents and modifiers of monogenic disorders. We further discuss how advances in functional genomics are now providing new opportunity to assess the impact of rare non-coding variants and their role in disease.

Keywords: rare variant, genetics of gene expression, RNA-sequencing, Mendelian disorders, eQTLs, allele-specific expression

INTRODUCTION

In the next few years, hundreds of thousands of genomes will be sequenced, exposing an unprecedented wealth of genetic information. As each new genetic variant offers a potential window into an individual's past, present, and future by providing insight into ancestry, traits, and disease risk, a major challenge will be to connect genetic variants to their functional consequences. Essential to addressing this challenge are high-throughput assays which connect genetic variants to molecular and cellular phenotypes. For instance, by assessing the association of genetic variants with the expression level of nearby genes (**Box 1**), a broad spectrum of causal variants, while potentially still unobserved, can be localized and their functional impact ascertained. At the moment, this experimental setting provides us with the most comprehensive system to identify regulatory variants and model the functional spectrum of human variation. Furthermore, genome-wide association studies (GWAS) have increasingly utilized such information to connect disease-predisposing variants to genes (Charlesworth et al., 2009; Montgomery and Dermitzakis, 2011). The basic principle behind such analyses has been to identify if a disease-predisposing variant is also associated to the expression level of a nearby gene – thereby providing a mechanistic hypothesis for disease etiology. In this area, multiple methods have been developed to assess the relationship between genetic effects on gene expression and disease by assessing the sharing of association (Zhu et al., 2007; Emilsson et al., 2008; Schadt et al., 2008; Nica et al., 2010). However, methods based on association do not well address the functional impact of rare, private, and *de novo* variants that dominate the site frequency spectrum in human populations. The extent of which was highlighted by two recent exome-sequencing that uncovered as much as 95% of protein-coding variation is rare (MAF $\leq$ 5%) and among these rare variants are the majority of sites predicted to be deleterious (Nelson et al., 2012; Tennessen et al., 2012). However, both studies focused on protein-coding variants where the genetic code greatly facilitates prediction of functional impact and not in the non-coding regions of the genome where considerable trait-predisposing variation is expected to reside (Hindorff et al., 2009). A considerable future challenge will be to determine the functional impact of a deluge of rare non-coding variants – as either causative agents or modifiers of traits. In this review, we first discuss the impact of non-coding variants as causal and modifying agents of monogenic disorders and then describe how advances in functional genomics can facilitate discovery and interpretation of rare, private, and *de novo* variants underlying a complete spectrum of human traits and diseases.

IMPORTANCE AND CHALLENGES IN IDENTIFYING NON-CODING VARIANTS

Advances in characterizing genome function have highlighted that, per base, non-coding sequence is at least as important for biochemical function as coding sequence (Birney et al., 2007; Ward and Kellis, 2012a). However, there is marked difference in the number of disease-predisposing variants that have been identified in the coding versus non-coding genome; for instance, the Human Gene Mutation database lists only a few percent of previously identified variants as being non-coding and regulatory (Stenson et al., 2009). This difference is attributable to how each class of variant is identified. For coding variants, the genetic code and data regarding protein structure and function facilitate predictions of causality. One can immediately identify if a variant changes an amino acid, creates a premature stop codon, alters hydrophobicity, disrupts a canonical splice junction, or impacts the structure of a protein domain. Furthermore, historically high

Box 1 | Statistical techniques for detecting genetic effects on gene expression

Expression quantitative trait loci (eQTL) mapping requires three important steps:

Data normalization:
Normalization is critical to mitigating the influence of technical artifacts or "batch effects" (Johnson et al., 2007; Williams et al., 2007) which can lead to spurious eQTL findings (Akey et al., 2007; Breitling et al., 2008). For microarray data, there are numerous normalization techniques to cope with batch effects (Irizarry et al., 2003; Johnson et al., 2007; Leek and Storey, 2007; Kang et al., 2008). For RNA-seq data, read counts can be transformed to RPKM (Reads Per Kilobase of transcript per Million mapped reads) (Mortazavi et al., 2008) or FPKM (Fragments Per Kilobase of transcript per Million mapped reads) (Trapnell et al., 2012) to account for differences in quantification due to transcript length and sequencing library depth. However, the utility of this approach is largely beneficial when comparing quantification between genes. As eQTL studies compare gene or transcript expression across samples, variability due to differences in transcript length can generally be ignored. However, in addition to library depth, the influence of highly expressed genes across samples can be profound – for instance, given equal numbers of reads sequenced, if a single gene (or small subset of genes) accounts for 10% of all reads in one sample and 5% of the reads in another, many other, unrelated genes may appear differentially expressed between these samples because reads not sequenced for one gene will yield extra reads sequenced for other genes. Such effects can be mitigated directly by inspection and regression or by using PCA-based normalization methods to correct out hidden factors (Pickrell et al., 2010). However, such normalization may be effective for dealing with global sequencing biases, individual genes can still be significantly influenced by read mapping errors (Marioni et al., 2008). Fundamental to addressing this issue are specific and sensitive read mapping algorithms for RNA-Seq data and an appropriate selection of an objective function for evaluating their performance in eQTL discovery. Here, number of eQTL discoveries is not sufficient as mappers which are highly influenced by variation may be enriched in false discoveries – advanced approaches account for alternative mappings and/or gain support from sequencing data across the expressed loci.

Association and linkage analysis:
In the presence of genetic marker data, different statistical tests are used to identify eQTL *within* populations and/or families (approaches for dealing with admixed or multi-population study designs are not covered in this review). Generally, the expression trait is assumed to be additive where each allele contributes equivalently. For population data, the most commonly used approach is a single-marker to single-trait *linear* or *non-parametric* regression; this analysis can be performed by statistical packages like R or statistical genetics tools such as PLINK (Purcell et al., 2007). For family data, linkage approaches based on inferring identity-by-descent (IBD) or linkage maps among family members are applied. Such a test can be conducted among all sib-pairs (Haseman and Elston, 1972) or all available family members (Amos, 1994). Many linkage analysis tools such as MERLIN (Abecasis et al., 2002) and SOLAR (Almasy and Blangero, 1998) can perform such analysis. For mixed designs, where data are composed of many nuclear families, transmission disequilibrium tests can be applied using tools like QTDT (Abecasis et al., 2000). However, each study design and associated statistical test has different advantages and limitations – association tests are powerful for detecting and localizing eQTL for common variants whereas linkage analyses cannot well localize variants but can identify intervals harboring rare eQTL.

Despite the availability of tools and simple statistical methods that support eQTL discovery, there remain important caveats in their utilization. For instance, considerable statistical power advantages can be achieved by restricting tests to variants nearby a gene (also termed *cis*-eQTL discovery). However, the size of the selected interval and the number of genes tested can be as influential as differences in sample size and number of variants tested. Complementary approaches have aimed to increase statistical power by integrating information from genes sharing regulatory pathways (Schadt et al., 2003). Here, the principal is that *bonafide* variant effects will propagate to downstream targets. For genome-wide (or *trans*-eQTL) analyses, many of the same limitations as present in GWA studies are equally applicable – common covariates such as ethnicity, relatedness, age and sex are also confounding factors in eQTL studies and should be appropriately accounted.

Though eQTL mapping is methodologically similar to GWA or family based linkage analysis, it can also be computationally more challenging when huge number of genes and variants need to be tested. Some newly developed computational tools are specifically optimized for fast eQTL mapping such as Matrix eQTL (Shabalin, 2012) and fastMap (Gatti et al., 2009).

Multiple testing correction:
Given the large numbers of tests and a typically small sample size, multiple testing correction is critical for eQTL mapping. Commonly used methods are Bonferroni correction, FDR/q-value (Hochberg and Benjamini, 1995; Storey and Tibshirani, 2003) and permutation-base *p*-value adjustment (Churchill and Doerge, 1994). A general practice of permutation is to generate a *p*-value distribution by randomly shuffling phenotype labels (typical for adjusting *cis p*-value) or by generating synthetic genotypes and phenotypes (typical for adjusting *trans p*-value). Using a permutation approach, a false-discovery rate can be estimated.

costs of sequencing have made it reasonable to focus exclusively on variant detection in coding sequences where gene structure is known. In contrast, for regulatory variants the size of the target region to survey is poorly defined. For example, it is not clear whether to screen a single kilobase from the transcription start site or 100 kb or whether to also include intronic and downstream sequence of a gene or even whether to assume the effect is not *cis*-acting at all and is located elsewhere in the genome. Complicating matters further is that given our incomplete understanding of regulatory architecture, even if a segregating non-coding variant is found, it has generally been difficult without further molecular biology-based experimentation to differentiate it from other variants and demonstrate its impact (Montgomery, 2009).

RARE NON-CODING VARIANTS UNDERLYING DISEASE

Despite the challenges of detecting causal regulatory variants, several studies have implicated them as the principal drivers of monogenic disorders. One of the earliest examples of a regulatory variant associated with a monogenic disorder was discovered in an individual affected with β-thalassemia (Poncz et al., 1982). The affected individual was identified to have an A–C transversion in a TATA-box element. Since this discovery, multiple regulatory variants associated with thalassemia have been described (Giardine et al., 2007). However, it is frequently the case that segregating non-coding variants lack differentiating genomic annotation and have required functional assays to support their effect. Such assays have generally included a combination of protein-expression, competition, electrophoretic mobility shift, and reporter gene assays. While such techniques are labor intensive, they have been critical to furthering the identification of non-coding variants underlying monogenic disorders. For instance, in Bernard–Soulier syndrome, an affected individual was identified with a G–C transversion within a predicted GATA-box motif (Ludlow et al., 1996). Subsequent functional assays supported the role of this variant by demonstrating *in vitro* an 84% reduction in promoter activity for the C allele and significant reduction in GATA-1 occupancy. Another similar example was described where a psoriasis-linked variant was identified through sequencing and predicted to influence a RUNX1 binding site. Subsequent functional assays confirmed to that the variant altered RUNX1 binding activity and reporter gene expression when RUNX1 and its coactivator CBFβ are present (Helms et al., 2003). We have further reported over 100 such variants where predicted regulatory polymorphisms have been assayed using a combination of electrophoretic mobility shift and reporter gene assays (Griffith et al., 2008). Despite this evidence however the *in vivo* activity and pathological cell types are rarely tested. Furthermore, the lack of availability of robust genome-wide assays for regulatory variation has been a challenge that is only more recently being addressed by advances in sequencing.

NON-CODING VARIANTS AS DISEASE MODIFIERS

Genetic modifiers can dramatically alter the penetrance of pathogenic mutations or influence the expressivity of disease. Modifiers need not be rare when involved in rare diseases but here we focus on their effect in monogenic disorders where we know the impact of modifiers is considerably influential. In monogenic diseases such as CPS1 deficiency, pathogenic and non-coding variant interaction can manifest strikingly different morbidities and mortalities ranging from asymptomatic to perinatal death (Klaus et al., 2009). This phenomenon is further well characterized by the phenotypic diversity diseases such as sickle cell anemia and cystic fibrosis. For instance, in sickle cell anemia, affected individuals can be protected by a higher level of HbF expression which substitutes function of a defected HBB gene (Steinberg and Adewoye, 2006). HbF is a gene coding fetal hemoglobin which is expressed only in residual levels in adults after a developmental switch to the adult hemoglobin gene HBB, however, adult HbF expression level is highly variable and genetically determined (Thein and Menzel, 2009). In cystic fibrosis, a genome-wide association study within affected individuals identified *IFRD1* as modulating disease phenotype by influencing pathogen defense and inflammation. Beyond modifier gene effects as in both sickle cell anemia and cystic fibrosis, pathogenic mutations and their associated disease phenotypes have been demonstrated to be modulated by multiple different types of modifier effects including variable expression of the primary disease gene itself as in thalassemia (Marzo et al., 2010; Sankaran et al., 2010) and allelic modulation of expression of defected and normal copies of a gene as in erythropoietic protoporphyria (Gouya et al., 1996). **Table 1** lists several well studied Mendelian diseases and an example of their modifier effects.

More recently, genome-wide surveys of allelic expression have highlighted the abundance of interaction between protein-coding and non-coding variation. These studies have estimated that as high as 20% of non-synonymous variants and 30% of genes have allelic-specific expression (ASE) (Dimas et al., 2008; Ge et al., 2009; Lappalainen et al., 2011; Montgomery et al., 2011). Through RNA-sequencing, allelic expression of protein-coding variants can be routinely assessed and offers the potential to survey this type of interaction for rare or *de novo* variants without requiring variant phasing (Pastinen, 2010). When the correct pathological tissue is interrogated, it is expected that future studies which integrate allelic expression data for their pathogenic mutations will routinely have improved capacity for interrogating the interaction between protein-coding and non-coding variation and predicting disease expressivity.

ASSIGNING FUNCTION TO RARE NON-CODING VARIANTS

LINKAGE ANALYSIS FOR GENE EXPRESSION

Linkage analysis is used in family based studies to detect *cis*- and *trans*-acting variants affecting genes expression. Gene expression level can be treated as a quantitative trait and analyzed by a general purpose linkage tool such as FBAT (Laird et al., 2000), GENEHUNTER TDT (Kruglyak et al., 1996), HRRR/HHRR (Terwilliger, 1995), LAMP (Li et al., 2005), MENDEL (Lange et al., 2005), PLINK (Purcell et al., 2007), PSEUDOMARKER (Hiekkalinna et al., 2011), QTDT (Abecasis et al., 2000), TRANSMIT (Clayton, 1999), UNPHASED (Dudbridge, 2008), and MERLIN (Abecasis et al., 2002). Many of these tools further model association in the presence of linkage and estimate the within family effect to the between family effects. Of these, the QTDT test has been well utilized in eQTL studies combining families and unrelated individuals (Cheung et al., 2010). A recent survey of the above methods has demonstrated improved power in the PSEUDOMARKER method relative to QTDT for which the authors account its ability to model true relationships between pedigrees, complete data usage and estimation of recombination, and allele frequencies from available data (Hiekkalinna et al., 2012). Using such methods, family studies have investigated the landscape of regulatory variants in different cell types such as lymphoblastoid cell lines (Schadt et al., 2003; Monks et al., 2004; Morley et al., 2004; Dixon et al., 2007; Cheung et al., 2010), lymphocytes (Göring et al., 2007), and adipose tissues (Emilsson et al., 2008). These studies have identified chromosomal regions co-segregating with transcriptome features, and most of them discover many eQTLs proximal to the genes they regulate. The reason why distal or *trans*-eQTLs are less detectable is in part due to a combination of smaller effect sizes and statistical power limitation by family size.

Table 1 | Mendelian disorders modified by gene expression.

Mendelian disorder with wide phenotypic diversity	Primary disease gene	Example of genetic modifier	Modifier effect	Reference
Cystic fibrosis	CFTR	IFRD1	Regulation of neutrophil effector function	Gu et al. (2009)
Sickle cell anemia	HBB	HbF	Substitutes function of HBB	Steinberg and Adewoye (2006)
Thalassemia	HBA, HBB	Promoter variant	Changes levels of HBA, HBB expressions	Weatherall (2001)
Hemochromatosis	HFE	TFR2	Co-modulator of hepcidin	Camaschella (2005)
Familial Hypercholesterolemia	LDLR	TNFRSF1B	Reduces shedding of the TNFRSF1B receptor	Geurts et al. (2000)
Hereditary deafness	DFNB26	DFNM1	Suppress DFNB26	Riazuddin et al. (2000)
Retinitis pigmentosa	RPGR	IQCB1	Interaction with RPGR	Fahim et al. (2011)
Familial Mediterranean fever	MEFV	MICA	MICA behaves as a stress-inducible self-antigen	Touitou et al. (2001)
Asthma drug response	ADRB2	Promoter variant	Alters ADRB2 receptor expression	Drysdale et al. (2000)
Gaucher disease	GBA	SCARB2	Causes extracellular excretion of GCase	Velayati et al. (2011)
Adrenoleukodystrophy	ABCD1	SOD2	Modulates the response of neurons to oxidative damage	Brose et al. (2012)
Alpha 1-antitrypsin deficiency	A1AT	NOS3	Regulates vascular tone	DeMeo (2004)
Wilson disease	ATP7B	PRNP	Produces prion protein also involved in transporting copper	Weiss et al. (2010)
Hereditary pancreatitis	PRSS1	SPINK1	Serine protease inhibitor	Weiss et al. (2003)
Polycystic kidney disease	PKD1, PKD2	ACE	Increase angiotensin II levels	Devuyst (2003)
Erythropoietic protoporphyria	FECH	Intronic variant	Reduces FECH activity	Gouya et al. (2006)

Despite restricted resolution in linkage analysis, family studies are especially useful for detecting potentially strong effect of rare variants. Deep sequencing of human exomes has detected an abundance of such rare variants (Nelson et al., 2012; Tennessen et al., 2012) in protein-coding sequences, however the extent to which rare variants have strong regulatory effects is only beginning to emerge. We recently reported increases in fold change for rare eQTLs when stratified by their derived allele frequency (Lappalainen et al., 2011). Existing linkage studies of gene expression are all based on families genotyped from SNP microarrays, yet to connect a discovered eQTL to a causal variant requires the use of deep sequencing data, which is also a necessary step for understanding the functional consequences of rarer variants.

RNA-SEQUENCING/ChIP-SEQUENCING

Compared with microarrays, RNA-sequencing can potentially cover gene activity over the whole transcriptome while providing higher resolution of transcript complexity (**Figure 1**). Two studies utilizing this new technology have discovered hundreds of genetic effects on gene expression in European and African populations (Montgomery et al., 2010; Pickrell et al., 2010). Furthermore, utilizing RNA-sequencing reads spanning splice junctions, an indicator of alternative splicing events, studies have also shown wide spread splicing polymorphisms between individuals (Lalonde et al., 2011). However RNA-sequencing has its own technical biases (Mortazavi et al., 2008; Labaj et al., 2011), highly expressed genes may consume majority of reads which leaves many genes below quantifiable threshold, mapping artifacts may introduce false positive associations and technical biases like library construction and PCR-based amplification may further distort true allelic ratios. From the perspective of genome mapping, treatment of RNA reads is also substantially harder than DNA as reads from mRNA transcripts are gapped by introns; however, computational methods specifically designed for RNA-Seq data such as Tophat/Cufflinks tools (Trapnell et al., 2012) are increasingly available to support mapping and quantification of RNA reads for gene expression and alternative splicing analysis.

As gene expression is closely related to transcription factor binding, genetic studies of transcription factor binding offer insight into the functional of non-coding variants. Here, chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) can provide a comprehensive survey of transcription factor binding sites across the genome. ChIP-seq studies comparing related and unrelated individuals (McDaniell et al., 2010) have demonstrated heritability of chromatin structure and transcription factor binding, which has also been shown

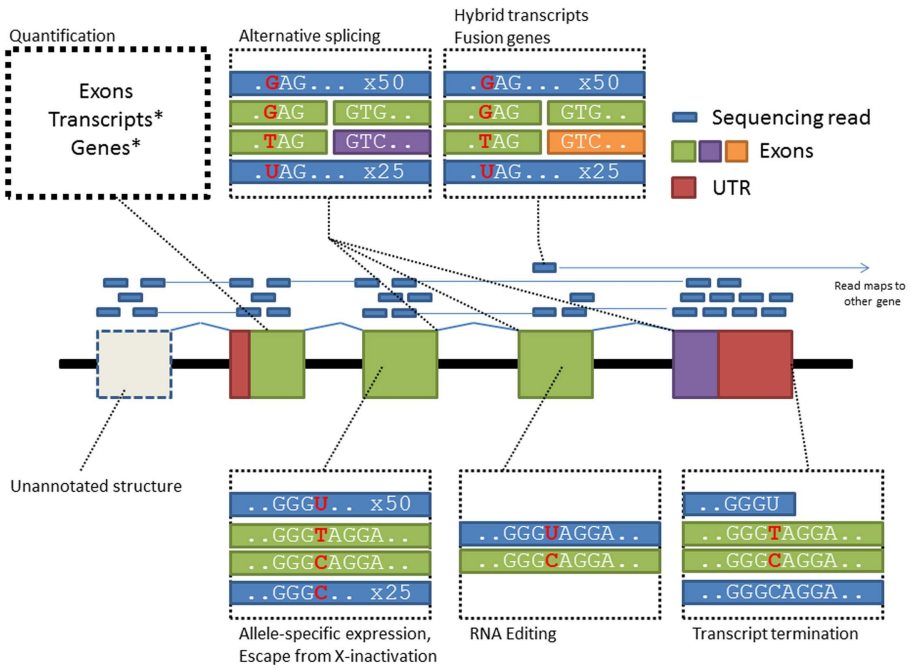

FIGURE 1 | RNA-Seq supports the characterization of diverse transcriptome features providing increased ability for linking trait-predisposing variation to their mode of impact. Each box in this figure highlights a specific-type of biological data that can be assessed from RNA-Seq data.

as a result of underlying genetic variation. Family information can be further utilized to distinguish the effects of rare variants and identify difference in binding of transcription factors specific to paternal and maternal alleles. One ChIP-seq study used phased diploid genomes from a family trio (Rozowsky et al., 2011) and by linking ASE to allele-specific binding of transcription factors connected expression of a gene to transcription factor binding on the same chromosome.

ALLELE-SPECIFIC EXPRESSION

Allele-specific expression has also been shown to be a heritable trait under genetic control through family studies (Yan et al., 2002) and studies of monozygotic twins (Cheung et al., 2008), with 0.47–0.98 correlation between monozygote twins. Allelic expression is a more sensitive indicator of *cis*-acting effects (Pastinen et al., 2006; Ge et al., 2009) and can be used as supporting evidence for the presence of *cis*-regulatory variants near a gene (Montgomery et al., 2011). A specific advantage of the availability of functional genomic studies like RNA-Seq and ChIP-Seq using high-throughput sequencing has been that allelic effects can be ascertained *en masse* by assessing biases in sequencing reads over heterozygous positions (the null being that both alleles are equally present). However, robust calls of ASE typically require heterozygous sites to be identified through genotyping or DNA sequencing. Genotypes may be directly observed from RNA data, however due to systematic variation of read depth of RNA (since transcripts are expressed at different levels and have different sizes) and reference genome mapping biases (due to unobserved variants or homologous sequences) such approaches are error prone (Degner et al., 2009; Heap et al., 2010) and may further fail at sites which are monoallelically expressing and appear homozygous. Furthermore, as read depth is only evaluated at a single variant site, there can be considerable variation due to random sampling effect, which overshadows most allelic effects with small fold changes. We have previously used a modified binomial test to assess significance of skewed expression at a site (Montgomery et al., 2010). However, the power of the test to distinguish true ASE from random sampling is determined by read depth at that site. Due to variation in expression levels of different genes, only a small proportion of sites may reach required confidence level. Targeted approaches to assessing ASE have been reported which combining capture technology over previously identified heterozygous positions and may ultimately offer more uniform assessments of allelic effects (Zhang et al., 2009). The advantage of high-quality allelic expression information will provide extra information to aid in mapping *cis*-eQTL (Sun, 2012) and causal regulatory variants (Montgomery et al., 2011). For protein-coding variation implicated in disease, it will further demarcate the bounds of haploinsufficiency such that disease risk may be stratified by the levels of functional protein product an individual produces.

CONNECTING EXPRESSION QUANTITATIVE TRAIL LOCI TO HUMAN DISEASE

Most early attempts to connect expression to phenotypic traits in families have involved using standard Pearson correlation between the expression and trait measurements. However, despite the statistical simplicity of this methodology, it can inflate significance when there are few families and the trait and expression variance has a large between family component, and can reduce significance if there is a large within family component and no correlation

between families. To address this, expression trait concordance tests which account for family structure have been reported (Kraft et al., 2003; Lu et al., 2004).

When combining genetic data to search for concordance of expression and trait association, an advantage is that the component of expression due to technical or environmental perturbation can be better controlled. We have previously reported methods that connect genetic effects on gene expression in unrelated samples to trait values by assessing if the genetic perturbation on expression is similar to the distribution of association scores for a trait (Nica et al., 2010). We have further developed a method to assess if there is an excess of causal regulatory variants of any frequency underlying a disease-associated variant (Conde et al., 2013). Here, our method tests if there are more ASE effects for heterozygotes of the disease-associated variant than homozygotes under the assumption that there may be one or more variants which stratify the risk and protective-associated alleles. Such approaches are now further complemented by the increasing amount of non-coding annotation from ChIP-Seq data and associated methods have been developed to determine the context and impact of trait-associated variants on epigenetic states (Boyle et al., 2012; Ward and Kellis, 2012b).

FUTURE PERSPECTIVES

As genome studies have been identifying large numbers of rare variants, it is expected that new methods and data will be required to uncover the impact of these variants and their involvement in diseases and traits. One of the most promising future technologies which will aid in interpreting the effect of rare regulatory variants in different tissues and developmental stages will be through the generation of induced pluripotent cells where gene expression can be assayed in multiple stages and differentiated cell types. Complementing this will be further advances in sequencing methods which provide phasing information – these advances will position rare and *de novo* mutations on the correct haplotype background and aid in investigations of genetic interaction. Furthermore, advances in ChIP-Seq and RNA-Seq will better aid in characterization of genetic effects on transcription factor binding and isoform expression and will ultimately unlock more complex functional interactions that underlie the etiology of diverse traits.

REFERENCES

Abecasis, G. R., Cardon, L. R., and Cookson, W. O. (2000). A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* 66, 279–292.

Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* 30, 97–101.

Akey, J. M., Biswas, S., Leek, J. T., and Storey, J. D. (2007). On the design and analysis of gene expression studies in human populations. *Nat. Genet.* 39, 807–808; author reply 808–809.

Almasy, L., and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* 62, 1198–1211.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* 54, 535–543.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797.

Breitling, R., Li, Y., Tesson, B. M., Fu, J., Wu, C., Wiltshire, T., et al. (2008). Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.* 4:e1000232. doi:10.1371/journal.pgen.1000232

Brose, R. D., Avramopoulos, D., and Smith, K. D. (2012). SOD2 as a potential modifier of X-linked adrenoleukodystrophy clinical phenotypes. *J. Neurol.* 259, 1440–1447.

Camaschella, C. (2005). Understanding iron homeostasis through genetic analysis of hemochromatosis and related disorders. *Blood* 106, 3710–3717.

Charlesworth, J. C., Peralta, J. M., Drigalenko, E., Göring, H. H., Almasy, L., Dyer, T. D., et al. (2009). Toward the identification of causal genes in complex diseases: a gene-centric joint test of significance combining genomic and transcriptomic data. *BMC Proc.* 3(Suppl. 7), S92.

Cheung, V. G., Bruzel, A., Burdick, J. T., Morley, M., Devlin, J. L., and Spielman, R. S. (2008). Monozygotic twins reveal germline contribution to allelic expression differences. *Am. J. Hum. Genet.* 82, 1357–1360.

Cheung, V. G., Nayak, R. R., Wang, I. X., Elwyn, S., Cousins, S. M., Morley, M., et al. (2010). Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol.* 8:e1000480. doi:10.1371/journal.pbio.1000480

Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963–971.

Clayton, D. (1999). A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am. J. Hum. Genet.* 65, 1170–1177.

Conde, L., Bracci, P. M., Richardson, R., Montgomery, S. B., and Skibola, C. F. (2013). Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma. *Am. J. Hum. Genet.* 92, 126–130.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., et al. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 25, 3207–3212.

DeMeo, D. L. (2004). α1-Antitrypsin deficiency 2: genetic aspects of α1-antitrypsin deficiency: phenotypes and genetic modifiers of emphysema risk. *Thorax* 59, 259–264.

Devuyst, O. (2003). Autosomal dominant polycystic kidney disease: modifier genes and endothelial dysfunction. *Nephrol. Dial. Transplant.* 18, 2211–2215.

Dimas, A. S., Stranger, B. E., Beazley, C., Finn, R. D., Ingle, C. E., Forrest, M. S., et al. (2008). Modifier effects between regulatory and protein-coding variation. *PLoS Genet.* 4:e1000244. doi:10.1371/journal.pgen.1000244

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C., et al. (2007). A genome-wide association study of global gene expression. *Nat. Genet.* 39, 1202–1207.

Drysdale, C. M., McGraw, D. W., Stack, C. B., Stephens, J. C., Judson, R. S., Nandabalan, K., et al. (2000). Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10483–10488.

Dudbridge, F. (2008). Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum. Hered.* 66, 87–98.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., et al. (2008). Genetics of gene expression and its effect on disease. *Nature* 452, 423–428.

Fahim, A. T., Bowne, S. J., Sullivan, L. S., Webb, K. D., Williams, J. T., Wheaton, D. K., et al. (2011). Allelic heterogeneity and genetic modifier loci contribute to clinical variation in males with X-linked retinitis pigmentosa due to RPGR mutations. *PLoS ONE* 6:e23021. doi:10.1371/journal.pone.0023021

Gatti, D. M., Shabalin, A. A., Lam, T. C., Wright, F. A., Rusyn, I., and Nobel, A. B. (2009). FastMap: fast eQTL mapping in homozygous populations. *Bioinformatics* 25, 482–489.

Ge, B., Pokholok, D. K., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D. J., et al. (2009). Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.* 41, 1216–1222.

Geurts, J. M., Janssen, R. G., van Greevenbroek, M. M., van der Kallen, C. J., Cantor, R. M., Bu, X., et al. (2000). Identification of TNFRSF1B as a novel modifier gene in familial combined hyperlipidemia. *Hum. Mol. Genet.* 9, 2067–2074.

Giardine, B., van, B. aalS., Kaimakis, P., Riemer, C., Miller, W., Samara, M., et al. (2007). HbVar database of human hemoglobin variants and

thalassemia mutations: 2007 update. *Hum. Mutat.* 28, 206.

Göring, H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., Charlesworth, J., Cole, S. A., et al. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* 39, 1208–1216.

Gouya, L., Deybach, J. C., Lamoril, J., Da Silva, V., Beaumont, C., Grandchamp, B., et al. (1996). Modulation of the phenotype in dominant erythropoietic protoporphyria by a low expression of the normal ferrochelatase allele. *Am. J. Hum. Genet.* 58, 292–299.

Gouya, L., Martin-Schmitt, C., Robreau, A. M., Austerlitz, F., Da Silva, V., Brun, P., et al. (2006). Contribution of a common single-nucleotide polymorphism to the genetic predisposition for erythropoietic protoporphyria. *Am. J. Hum. Genet.* 78, 2–14.

Griffith, O. L., Montgomery, S. B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., et al. (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.* 36, D107–D113.

Gu, Y., Harley, I. T., Henderson, L. B., Aronow, B. J., Vietor, I., Huber, L. A., et al. (2009). Identification of IFRD1 as a modifier gene for cystic fibrosis lung disease. *Nature* 458, 1039–1042.

Haseman, J. K., and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* 2, 3–19.

Heap, G. A., Yang, J. H., Downes, K., Healy, B. C., Hunt, K. A., Bockett, N., et al. (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.* 19, 122–134.

Helms, C., Cao, L., Krueger, J. G., Wijsman, E. M., Chamian, F., Gordon, D., et al. (2003). A putative RUNX1 binding site variant between SLC9A3R1 and NAT9 is associated with susceptibility to psoriasis. *Nat. Genet.* 35, 349–356.

Hiekkalinna, T., Göring, H. H., Lambert, B., Weiss, K. M., Norrgrann, P., Schäffer, A. A., et al. (2012). On the statistical properties of family-based association tests in datasets containing both pedigrees and unrelated case-control samples. *Eur. J. Hum. Genet.* 20, 217–223.

Hiekkalinna, T., Schäffer, A. A., Lambert, B., Norrgrann, P., Göring, H. H., and Terwilliger, J. D. (2011). PSEUDOMARKER: a powerful program for joint linkage and/or linkage disequilibrium analysis on mixtures of singletons and related individuals. *Hum. Hered.* 71, 256–266.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367.

Hochberg, Y., and Benjamini, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* 57, 289–300.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127.

Kang, H. M., Ye, C., and Eskin, E. (2008). Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180, 1909–1925.

Klaus, V., Vermeulen, T., Minassian, B., Israelian, N., Engel, K., Lund, A. M., et al. (2009). Highly variable clinical phenotype of carbamylphosphate synthetase 1 deficiency in one family: an effect of allelic variation in gene expression? *Clin. Genet.* 76, 263–269.

Kraft, P., Schadt, E., Aten, J., and Horvath, S. (2003). A family-based test for correlation between gene expression and trait values. *Am. J. Hum. Genet.* 72, 1323–1330.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* 58, 1347–1363.

Labaj, P. P., Leparc, G. G., Linggi, B. E., Markillie, L. M., Wiley, H. S., and Kreil, D. P. (2011). Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics* 27, i383–i391.

Laird, N. M., Horvath, S., and Xu, X. (2000). Implementing a unified approach to family-based tests of association. *Genet. Epidemiol.* 19(Suppl. 1), S36–S42.

Lalonde, E., Ha, K. C., Wang, Z., Bemmo, A., Kleinman, C. L., Kwan, T., et al. (2011). RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res.* 21, 545–554.

Lange, K., Sinsheimer, J. S., and Sobel, E. (2005). Association testing with Mendel. *Genet. Epidemiol.* 29, 36–50.

Lappalainen, T., Montgomery, S. B., Nica, A. C., and Dermitzakis, E. T. (2011). Epistatic selection between coding and regulatory variation in human evolution and disease. *Am. J. Hum. Genet.* 89, 459–463.

Leek, J. T., and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3:e161. doi:10.1371/journal.pgen.0030161

Li, M., Boehnke, M., and Abecasis, G. R. (2005). Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. *Am. J. Hum. Genet.* 76, 934–949.

Lu, Y., Liu, P. Y., Liu, Y. J., Xu, F. H., and Deng, H. W. (2004). Quantifying the relationship between gene expressions and trait values in general pedigrees. *Genetics* 168, 2395–2405.

Ludlow, L. B., Schick, B. P., Budarf, M. L., Driscoll, D. A., Zackai, E. H., Cohen, A., et al. (1996). Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibbeta promoter resulting in the Bernard-Soulier syndrome. *J. Biol. Chem.* 271, 22076–22080.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18, 1509–1517.

Marzo, R., Calzolari, R., and Rund, D. (2010). "Modifier genes in the haemoglobinopathies," in *Encyclopedia of Life Sciences* (Chichester: John Wiley & Sons Ltd), 1–9.

McDaniell, R., Lee, B. K., Song, L., Liu, Z., Boyle, A. P., Erdos, M. R., et al. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* 328, 235–239.

Monks, S. A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., et al. (2004). Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* 75, 1094–1105.

Montgomery, S. (2009). Current computational methods for prioritizing candidate regulatory polymorphisms. *Methods Mol. Biol.* 569, 89–114.

Montgomery, S. B., and Dermitzakis, E. T. (2011). From expression QTLs to personalized transcriptomics. *Nat. Rev. Genet.* 12, 277–282.

Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M., and Dermitzakis, E. T. (2011). Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet.* 7:e1002144. doi:10.1371/journal.pgen.1002144

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., et al. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773–777.

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., et al. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* 430, 743–747.

Mortazavi, A., Williams, B., McCue, K., and Schaeffer, L. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 1–8.

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St Jean, P., Verzilli, C., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337, 100–104.

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., et al. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 6:e1000895. doi:10.1371/journal.pgen.1000895

Pastinen, T. (2010). Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.* 11, 533–538.

Pastinen, T., Ge, B., and Hudson, T. J. (2006). Influence of human genome polymorphism on gene expression. *Hum. Mol. Genet.* 15, R9–R16.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., et al. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–772.

Poncz, M., Ballantine, M., Solowiejczyk, D., Barak, I., Schwartz, E., and Surrey, S. (1982). beta-Thalassemia in a Kurdish Jew. Single base changes in the T-A-T-A box. *J. Biol. Chem.* 257, 5994–5996.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575.

Riazuddin, S., Castelein, C. M., Ahmed, Z. M., Lalwani, A. K., Mastroianni,

M. A., Naz, S., et al. (2000). Dominant modifier DFNM1 suppresses recessive deafness DFNB26. *Nat. Genet.* 26, 431–434.

Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., et al. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* 7, 522.

Sankaran, V. G., Lettre, G., Orkin, S. H., and Hirschhorn, J. N. (2010). Modifier genes in Mendelian disorders: the example of hemoglobin disorders. *Ann. N. Y. Acad. Sci.* 1214, 47–56.

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., et al. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* 6:e107. doi:10.1371/journal.pbio.0060107

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422, 297–302.

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28, 1353–1358.

Steinberg, M. H., and Adewoye, A. H. (2006). Modifier genes and sickle cell anemia. *Curr. Opin. Hematol.* 13, 131–136.

Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Thomas, N. S., et al. (2009). The human gene mutation database: 2008 update. *Genome Med.* 1, 13.

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445.

Sun, W. (2012). A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* 68, 1–11.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 64, 64–69.

Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.* 56, 777–787.

Thein, S. L., and Menzel, S. (2009). Discovering the genetics underlying foetal haemoglobin production in adults. *Br. J. Haematol.* 145, 455–467.

Touitou, I., Picot, M. C., Domingo, C., Notarnicola, C., Cattan, D., Demaille, J., et al. (2001). The MICA region determines the first modifier locus in familial Mediterranean fever. *Arthritis Rheum.* 44, 163–169.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578.

Velayati, A., DePaolo, J., Gupta, N., Choi, J. H., Moaven, N., Westbroek, W., et al. (2011). A mutation in SCARB2 is a modifier in Gaucher disease. *Hum. Mutat.* 32, 1232–1238.

Ward, L. D., and Kellis, M. (2012a). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337, 1675–1678.

Ward, L. D., and Kellis, M. (2012b). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–D934.

Weatherall, D. J. (2001). Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. *Nat. Rev. Genet.* 2, 245–255.

Weiss, F. U., Simon, P., Witt, H., Mayerle, J., Hlouschek, V., Zimmer, K. P., et al. (2003). SPINK1 mutations and phenotypic expression in patients with pancreatitis associated with trypsinogen mutations. *J. Med. Genet.* 40, e40.

Weiss, K. H., Runz, H., Noe, B., Gotthardt, D. N., Merle, U., Ferenci, P., et al. (2010). Genetic analysis of BIRC4/XIAP as a putative modifier gene of Wilson disease. *J. Inherit. Metab. Dis.* doi:10.1007/s10545-010-9123-5

Williams, R. B. H., Chan, E. K. F., Cowley, M. J., and Little, P. F. R. (2007). The influence of genetic variation on gene expression. *Genome Res.* 17, 1707–1716.

Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., and Kinzler, K. W. (2002). Allelic variation in human gene expression. *Science* 297, 1143.

Zhang, K., Li, J. B., Gao, Y., Egli, D., Xie, B., Deng, J., et al. (2009). Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Methods* 6, 613–618.

Zhu, J., Wiener, M. C., Zhang, C., Fridman, A., Minch, E., Lum, P. Y., et al. (2007). Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput. Biol.* 3:e69. doi:10.1371/journal.pcbi.0030069

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 September 2012; accepted: 09 April 2013; published online: 31 May 2013.

Citation: Li X and Montgomery SB (2013) Detection and impact of rare regulatory variants in human disease. Front. Genet. ***4****:67. doi: 10.3389/fgene.2013.00067*

This article was submitted to Frontiers in Statistical Genetics and Methodology, a specialty of Frontiers in Genetics.